Aliah University

Department of Mathematics and Statistics

"Project Title: Predictive Survival Analysis Using Machine Learning Algorithms"

Project Work By: SNEHADRITA DAS

Final year student of Aliah University,

Submitted for the partial fulfilment of Masters in Statistics

Supervised By:

Dr. MOUMITA CHATTERJEE, Assistant Professor,

Dept of Mathematics and Statistics, Aliah University

Acknowledgement

I would like to express my deepest gratitude and appreciation to my supervisor, ***Dr.Moumita Chatterjee*** in this project work : ***Predictive Survival Analysis using Machine Learning Algorithms***, for her unwavering support, guidance, and inspiration throughout my journey. Her invaluable contributions have played a pivotal role in shaping my knowledge and exposure while working in this project. Beyond the classroom, my supervisor has been a source of encouragement and a pillar of strength. Her unwavering belief in my abilities has empowered me to overcome and deal with the challenges in this study. Whether it was offering a listening ear, providing insightful advice, she has been an incredible mentor. In conclusion, I am truly fortunate to have had Dr.Moumita Chatterjee as my supervisor and mentor. Her unwavering support, guidance, and dedication have been instrumental and because of that I got to learn an incredible amount of impactful things. At last but not the least, I would like to thank my parents a great deal. Their support, love and care go beyond words.

Yours Thankfully,
Snehadrita Das

Introduction

A problem frequently faced by applied statisticians is the analysis of time to event data. Examples of such data arise in diverse fields such as medicine, biology, public health, epidemiology, engineering, economics and demography. While the statistical tools we shall present are applicable to all these disciplines our focus is on applications of the techniques to biology and medicine. Survival time can be defined broadly as the time to the occurrence of a given event. This event can be the development of a disease, response to a treatment,relapse, or death. Therefore, survival time can be tumor-free time, the time from the start of treatment to response, length of remission, and time to death. Survival data can include survival time, response to a given treatment, and patient characteristics related to response, survival, and the development of a disease. The study of survival data has focused on predicting the probability of response, survival, or mean lifetime, comparing the survival distributions of experimental animals or of human patients and the identification of risk and/or prognostic factors related to response, survival, and the development of a disease.

The analysis of survival experiments is complicated by issues of censoring, where an individual's life length is known to occur only in a certain period of time where individuals enter the study only if they survive a sufficient length of time or individuals are included in the study only if the event has occurred by a given date. The emphasis is on applying, exploring and comparing the survival analysis techniques in semi parametric and nonparametric setup.

The objective of this study is to compare the performances of Cox Proportional Hazard Regression(CoxPH) and Random Survival Forests (RSF) methods with a real data set related to Bone Marrow transplantation. Most popular of survival analyses is Cox regression analysis because it is a semiparametric method for investigating the effect of several variables upon the time a specified event takes to happen. Recently, random survival forests (RSF) (Ishwaran et al. (2008)) has been used for the analysis of survival data. It is an ensemble tree method for the analysis of right censored survival data.

The motivation of this study is to identify the most important factors influencing the success or failure of the transplantation procedure. Healthcare data are valuable, take a lot of money, time and man power to collect and hence we might not want to throw away variables unnecessarily. Hence the priority is to optimize the model accuracy given that we are able to retain most number of variables.

In this study we consider pediatric patients with several hematologic diseases: malignant disorders with acute leukemia, with chronic leukemia, with myelodysplastic syndrome, and nonmalignant cases. All patients were subject to the unmanipulated allogeneic unrelated donor hematopoietic stem cell transplantation.
The set contains 187 observations characterized by 33 attributes.

Material and Methodology

Material : Dataset

Created by Institute of Computer Science, Silesian University of Technology,Poland and Institute of Innovative Technologies EMAG, Poland and others, the dataset contains *177 observations characterized by 33 variables*. Among these, we have *9 continuous and 22 categorical variables*.

A quick look to some of the variables

- Stemcellsource - Source of hematopoietic stem cells
- IIIV - Development of acute graft versus host disease stage II or III or IV (Yes - 1, No - 0)
- Diseasegroup - Type of disease (malignant - 1, nonmalignant - 0)
- Relapse - Reoccurrence of the disease (No - 0, Yes - 1)
- CD34kgx10d6 - CD34+ cell dose per kg of recipient body weight (10ˆ6/kg)

Methodologies

Cox Proportional Hazard Regression

The Cox model is the most popular regression model for survival data. Its properties can be derived using the counting process techniques of Aalen. There are inference packages for it in almost every statistical package. It can be extended quite easily to time-dependent covariates and models for multistate models or models with random effects.

An important problem is to predict the distribution of the time to some event from a set of explanatory variables. Here, the interest is in predicting risk factors for the event of interest. Statistical strategies for prediction are similar to those utilized in ordinary regression. However,the details for regression techniques in survival studies are unique.
Let $h(t|\mathbf{x})$ be the hazard rate at time t for an individual, The basic model due to Cox (1972) is as follows:

$h(t|\mathbf{x}) = h_0(t)exp(\mathbf{b}'\mathbf{x})$

where $\mathbf{x}$ is the covariate vector $h_0(t)$ is an arbitrary baseline hazard rate and $\mathbf{b} = (\beta_1, \beta_2, ..., \beta_p)'$ is a parameter vector.
This is called a semiparametric model because a parametric form is assumed only for the covariate effect. The baseline hazard rate is treated nonparametrically. The coding of factors and their interaction effects follows the usual rules for linear models. An interaction between two or more factors may be examined by constructing new variables which are the product of the variables associated with the individual factors as is commonly done in other regression contexts.
The Cox model is often called a proportional hazards model because, if we look at two individuals with covariate vector values such as $\mathbf{x}_1$ and $\mathbf{x}_2$ then,the ratio of their hazards is,

$\frac{h(t|\mathbf{x}_1)}{h(t|\mathbf{x}_2)} = \frac{h_0(t)exp(\mathbf{b}'\mathbf{x}_1)}{h_0(t)exp(\mathbf{b}'\mathbf{x}_2)} = \frac{exp(\mathbf{b}'\mathbf{x}_1)}{exp(\mathbf{b}'\mathbf{x}_2)}$

which is a constant. So, the hazard rates are proportional. This quantity is called the relative risk (hazard ratio) of an individual with risk factor $textbf{x}_1$ having the event as compared to an individual with risk factor $textbf{x}_2$.

- ***Partial Likelihood***:A definition is given of partial likelihood generalizing the ideas of conditional and marginal likelihood. Applications include life tables and inference in stochastic processes. It is shown that the usual large-sample properties of maximum likelihood estimates and tests apply when partial likelihood is used.
- Merit of partial likelihood method : The method involves less assumptions and hence is more robust than full likelihood method.
- Demerit of partial likelihood method : The method is less powerful compared to a fully parametric model.

The regression vector $textbf{b}$ is estimated without making any assumptions about the functional form of the baseline hazard rate, by maximizing the marginal, partial or maximum likelihood functions which are obtained by considering the contribution to the hazard rate of the individual time to failure. The marginal likelihood method for estimating vector/3 is based on the marginal distribution of the rank statistics of the times to the failures.

Typically the goal of an investigation is to make an inference about $\mathbf{b}$. This uses the partial likelihood method for estimating the regression coefficients,

$L(\mathbf{b}) = \prod_{i=1}^{k} \frac{exp(\mathbf{b}'\mathbf{x}_{(i)})}{\sum_{l \in R(t_{(i)})} exp(\mathbf{b}'\mathbf{x}_{(l)})}$

where where the summation in the denominator is over all subjects in the risk set at time t_i is denoted by $R(t_{(i)})$,the product is over the k distinct ordered survival times, $\mathbf{x}_i$ denotes the e value of the covariate for the subject with ordered survival time t_i.
This is treated as a usual likelihood, and inference is carried out by usual means. It is of interest to note that the numerator of the likelihood depends only on information from the individual who experiences the event, whereas the denominator utilizes information about all individuals who have not yet experienced the event (including some individuals who will be censored later). As in regression models for other types of data, the covariate vector X can contain transformations and interactions of the risk factors. It should be noted that there is no constant term in the regression vector. The constant is absorbed in the baseline hazard.
The *Cox Regression has a couple of assumptions*, that is,
* The hazard rate of a system is the product of the baseline hazard which is dependent on time only and a function of covariates, independent if time. That implies that the ratio of hazards are independent of time.
* Another assumption is that, they specify a multiplicative relationship between the underlying hazard function and the log-linear function of the covariates.

Modification on PHM : Stratified Cox

If the proportional hazard assumption is violated for a variable, then, one approach to dealing with this problem is to stratify on this variable. Stratification fits a different baseline hazard function for each stratum, so that the form of the hazard function for different levels of this variable is not constrained by their hazards being proportional. It is assumed, however, that the proportional hazards model is appropriate within strata for the other covariates. Usually one assumes the same parameters for the other variables in each stratum. One simple approach to this problem is to fit a model with an indicator function. In the simplest approach, we define a time dependent covariate and we have a proportional hazards model with hazard rate

$h(t|Z) = h_0(t)exp(b_1 z_1)$ if $t \leq \tau$ and $h(t|Z) = h_0(t)exp((b_1 + b_2)z_1)$ if $t \geq \tau$, where $Z = z_1$ if $t \geq \tau$ and $Z = 0$ if $t \leq \tau$

If the value of the covariates considered in this model depends on time, the model can no longer be described as a proportional hazards model since as time varies so will the ratio between the hazard rates. Hence, the term nonproportional hazards model is more suitable. This situation can be used to test the nonproportionality of hazards by introducing a dummy time dependent covariate as suggested by Cox. The partial likelihood estimates of the parameters may be explained as the short and long term effects of the covariate. Although it is possible to form a two or more steps model, depending on how a covariate changes with time, a sufficient number of times to failure should be available for each step.

Diagnostics of Cox PHM : Schoenfeld and Martingale Residuals

Graphical methods are based on residuals and are often used as diagnostic tools.In multiple regression methods, residuals are referred to as the difference between the observed and the predicted values (based on the regression model) of the dependent variable.However, when censored observations are present and only a partial likelihood function is used in the proportional hazards model, the usual concept of residuals is not applicable.

- A transformed plot of the partial residuals suggested by Schoenfeld can also be used as an exploratory tool to detect the time varying effects of a covariate. The Schoenfeld residuals are calculated for all covariates for each individual experiencing an event at a given time. Those are the differences between that individual's covariate values at the event time and the corresponding risk-weighted average of covariate values among all those then at risk. The word "residual" thus makes sense, as it's the difference between an observed covariate value and what you might have expected based on all those at risk at that time.
- Martingale residual (Fleming and Harrington, 1991), is the difference between the estimated accumulated hazard based on the proportional hazards model and the indicator (1 if the event occurs and 0 for censored time point). It has a skewed distribution with mean zero. Martingale residuals can be used to assess the true functional form of a particular covariate. Martingale residuals can also be used to assess outliers in the data set whereby the survivor function predicts an event either too early or too late, however, it's often better to use the deviance residual for this.

Regularised Cox Regression : Cox-LASSO

L1(LASSO) and L2(Ridge) penalized estimation methods shrink the estimates of the regres sion coefficients towards zero relative to the maximum likelihood estimates. L1 penalty on regression coefficients,providing what we call the adaptive Lasso estimator. The method incorporates different penalties for different coefficients: unimportant variables receive larger penalties than important ones, so that important variables tend to be retained in the selection process, whereas unimportant variables are more likely to be dropped. The purpose of this shrinkage is to prevent overfit arising due to either collinearity of the covariates or high-dimensionality.
Applying an L2 penalty tends to result in all small but non-zero regression coefficients, whereas applying an L1 penalty tends to result in many regression coefficients shrunk exactly to zero and a few other regression coefficients with comparatively little shrinkage. The amount of shrinkage is determined by tuning parameters λ_1 (for LASSO) and λ_2 (for Ridge). A value of zero always means no shrinkage (= maximum likelihood estimation) and a value of infinity means infinite shrinkage (= setting all regression coefficients to zero).

It is difficult to say in advance which value of lambdal or lambda2 to use. Note that for small values of lambdal or lambda2 the algorithm be very slow, may fail to converge or may run into numerical problems, especially in high-dimensional data. When this happens, we increase the value of lambdal or lambda2. Traditionally the choice of such parameters are determined by cross validation but in real time setup, the choice could be subject as well.

Random Survival Forest

- ***Usual Random Forest***: In RF, randomization is introduced in two forms. First, a randomly drawn bootstrap sample of the data is used to grow a tree. Second, at each node of the tree, a randomly selected subset of variables (covariates) is chosen as candidate variables for splitting. Averaging over trees, in combination with the randomization used in growing a tree, enables RF to approximate rich classes of functions while maintaining low generalization error.Random Survival Forest is an ensemble tree method for analysis of right-censored survival data. Constructing ensembles from base learners, such as trees, can substantially improve prediction performance. RSF strictly adheres to the prescription laid out of Random Forest by Breiman (2001). It based on a splitting rule and bootstrap samples. In RSF, randomization is introduced in two forms. First, a randomly drawn bootstrap sample of the data is used for growing the tree. Second, the tree learner is grown by splitting nodes on randomly selected predictors. While at first glance Random Forest might seem an unusual procedure, considerable empirical evidence has shown it to be highly effective.
- ***How RSF is different from RF***: In right-censored survival settings, this comprises survival time and censoring status. Thus, the splitting criterion used in growing a tree must explicity involve survival time and censoring information. Tree node impurity, measuring effectiveness of a split in separating data, must measure separation by survival difference. Further, the predicted value for a terminal node in a tree, the resulting ensemble predicted value from the forest, and the measure of prediction accuracy must all properly incorporate survival information.

Algorithm :
* Draw B bootstrap samples from the original data. Note that each bootstrap sample excludes on average 37% of the data, called out-of-bag data (OOB data).
* Grow a survival tree for each bootstrap sample. At each node of the tree, randomly select p candidate variables. The node is split using the candidate variable that maximizes survival difference between daughter nodes.
* Grow the tree to full size under the constraint that a terminal node should have no less than $d0 > 0$ unique deaths.
* Calculate a CHF for each tree. Average to obtain the ensemble Cumulative hazard function (CHF).
* Using OOB data, calculate prediction error for the ensemble CHF.

Different Splitting Rules :
* A log-rank splitting rule (logrank) that splits nodes by maximization of the logrank test statistic.
* A log-rank score rule (logrankscore) that splits nodes using a standardized logrank statistic.
* A random log-rank splitting rule (logrankrandom). A random split is selected for each of the p candidate variables in a node, and the variable with maximum log-rank statistic (at its random split point) is used to split the node.

Performance of RF regression depended strongly on the censoring rate.

- ***Variable Importance (VIMP) :*** RSF can be used for variable selection as well, Large importance values indicate variables with predictive ability, whereas zero or negative values identify nonpredictive variables to be filtered. Under the C-index, one can interpret VIMP in terms of misclassification. VIMP measures the change in prediction error on a fresh test case if x were not available, given that the original forest was grown using x. Although, in practice, this often equals change in prediction error for a forest grown with and without x, conceptually the two quantities are different.
- ***Prediction Error:*** To estimate prediction error, we use Harrell's concordance index [Harrell et al. (1982)]. The C-index (concordance index) is related to the area under the ROC curve [Heagerty and Zheng (2005)]. It estimates the probability that, in a randomly selected pair of cases, the case that fails first had a worst predicted outcome. The interpretation of the C-index as a misclassification probability is attractive, and is one reason we use it for prediction error. Another attractive feature is that, unlike other measures of survival performance, the C-index does not depend on a single fixed time for evaluation. The C-index also specifically accounts for censoring.

Data Analysis

1. Preprocessing

1.1. Data Imbalance :

A classification data set with skewed class proportions is called imbalanced. Classes that make up a large proportion of the data set are called majority classes. Those that make up a smaller proportion are minority classes.Data imbalance usually reflects an unequal distribution of classes within a dataset.

1.2. Dealing with Data Imbalance:

- One way of dealing with this situation is to club the levels of the categorical variables to bring symmetry in the frequency for each class.Since only the number of levels are getting reduced, we lose no information.
- Also, too few data points for a class would result inconsistent estimate for the coefficient of that level in question. Hence by clubbing, we bring symmetry and consistent estimates of the parameters.Imbalanced data can cause models to struggle in correctly predicting the minority class. The model may have a tendency to predict the majority class more frequently, resulting in low recall or sensitivity for the minority class. This is particularly problematic when the minority class contains crucial or critical instances that require accurate identification.
- For our data, variables like, ***HLAgrI*** - The difference type in terms of cells or antigens between the donor and the recipient, had had 7 levels with highly imbalanced structure now after clubbing it has 2 levels, 0-no difference and 1-has difference. ***Allel*** - In how many allele there is difference beetwen the donor nad the recipient, had had 5 levels, now it has 2. 0-no difference and 1-difference.

1.3. Missing values:

- We had to filter 10 data points due to multiple missing observations in multiple columns, in this case we had no choice but to remove them from the set. Because imputing them in survival data setup was too complicated.

1.4. Data Splitting:

For implementing several machine learning algorithms, we have splitted the dataset into training and test set in approximately ratio of 8:2 that is 80% of the data is randomly sampled and labled as training set and the rest is test set.

Analysis

Analysis in survival setup is an important part of medical statistics, frequently used to define prognostic indices for mortality or recurrence of a disease, and to study the outcome of treatment.

2 Exploratory Data Analysis

Problems due to Multicollinearity

- It is usual to interpret a regression coefficient as measuring the change in the response variable when the corresponding predictor variable is increased by one unit and all other predictor variables are held constant. This interpretation may not be valid if there are strong linear relationships among the predictor variables. The condition of severe nonorthogonality (that is, the existence of strong linear relationships among the predictor variables) is also referred to as the ***problem of collinear data, collinearity, or multicollinearity***.
- In the presence of multicollinearity, the coefficients of the correlated variables become unstable and difficult to interpret. Multicollinearity inflates the standard errors of the coefficients, leading to wider confidence intervals and reduced precision. This, in turn, makes it more difficult to detect statistically significant relationships and can affect hypothesis testing and the reliability of the model.

Correlation among the Continuous variables

- As a part of EDA, we check the correlation among the continuous variables and find out that only ***Recipientage***(Age of the recipient of hematopoietic stem cells at the time of transplantation) and ***Rbodymass***(Body mass of the recipient of hematopoietic stem cells at the time of transplantation) are highly correlated, ***a magnitude of 0.9***.

2.2. Multiple Association among the Categorical variables

It is easy to calculate the association between two categorical variables at a time. Well established methods are available for that but to measure the association of a variable with all the others at the same time,there being no well established methodology available for computing the multiple association of more than two categorical variables, we had to do this in the other way around.

- We tried to measure the association or the variability of a variable explained with the help of deviance.The deviance is a concept in generalized linear models that measures the fitted generalized linear model with respect to a perfect model known as the saturated model.

- ***Null Deviance***: The null deviance is a generalization of the total sum of squares of the linear model. The null deviance shows how well the model predicts the response variable with only the intercept.
- ***Residual Deviance***: The residual deviance tells us how well the response variable can be predicted by a model with p predictor variables. The lower the value, the better the model is able to predict the value of the response variable.
- We fit ***logistic regressions*** for each variable having two levels as 0 and 1 as response and all the other categorical variables as predictors. We note down the null deviance and the residual deviance given by the fitted model.
- Then we calculate the quantity (Table 1) $\frac{D-D_p}{D}$ where D, D_p are null deviance and residual deviance of the fitted model for each variable, let's say, having p covariates, respectively. And take this as a metric to detect strong association among the categorical variables. We were able to come up with thresholds based on which we would suspect strong association and exclude the variable in question from our model for consistent estimates. If for a variable,the quantity obtained is greater than or equal to 0.3, we shall exclude that variable. This is based on a subjective choice. For example, setting Recipientage as response and other categorical variables as predictor had resulted in null deviance of 239.92 and residual deviance of 191.09, so the value of the metric is 0.2 wityh the AIC being 229.

- The argument behind doing such is that, too small value of the quantity, residual deviance, D_p indicates that the p covariates are explaining the variability of the response4 in question too well which might have been because of the strong relationship among them and if the residual deviance is small, the differenec of null and residual deviance would be large and hence a larger value of the difference of the null and residual deviance would be questionable.And we shall suspect the variable under study as collinear.

3. Initial CoxPH Model

We fit the Cox proportional hazard model to the data, excluding all the suspected collinear variable and when the model does not converge.

3.1. Reasons for CoxPH model to run out of iterations and not converging

- 1. Separation: When there are one or more predictor variables that perfectly predict the outcome variable, also known as complete separation, the maximum likelihood estimation may not exist, and the optimization algorithm may fail to converge.

- 2. Small sample size.

- 3. Highly correlated variables.

- 4. Non-proportional hazards.

Now we have checked for Multicollinearity and we do not have insufficient data points hence ***we suspect the presence of time dependent covariates*** and violation of the PHM assumption.

4. Diagnostic Plot : Schoenfeld Residual Plots

Based on the prior knowledge of the variables, we plot the schoenfeld residuals (Figure 2 and Figure 3) against the respective suspectable covariates and look for patterns, if patterens are present instead of randomly scattered residuals, we suspect time dependency of the covariate in question and hence violation of the proportional hazard assumption.
We deal with this situation by incorporating ***Stratified Cox***. And we refit the model after fixing for time dependency.
* (1).Survival curves for different strata must have hazard functions that are proportional over the time t
* (2).The relationship between the log hazard and each covariate is linear, which can be verified with residual plots.

5. Nested Model Comparisons

We compare among different nested models, that is, model with interaction term between the continuous variables and so on. But under no setup the interaction term was statistically significant.(Table 2)

6. Variable Selection : Step Regression

We perform the step regression for variable selection, after a certain number of steps, we end up with 10 statistically significant variables and an AIC value of the model to be 549.278. But in this usual procedure, we are loosing too many variables.

Table I

Response	Null Deviance(D)	Residual Deviance(Dp)	(D-Dp)/D	AIC
Recipientgender	239.92	191.09	0.20	229.09
Stemcellsource	186.68	163.22	0.12	201.22
Donorage35	242.88	229.82	0.05	267.82
IIIV	237.58	167.35	0.30	205.35
Gendermatch	157.88	107.94	0.31	145.94
DonorABO	189.17	162.31	0.14	200.31
RecipientABO	228	199.52	0.12	237.52
RecipientRh	147.72	128.43	0.13	166.43
ABOmatch	210.73	178.81	0.15	216.81
DonorCMV	236.71	213.58	0.097	251.58
Riskgroup	231.63	130.93	0.43	168.93
Txpostrelapse	124.87	57.495	0.53	95.495
Diseasegroup	161.1	114.14	0.29	152.14
HLAmatch	245.37	~0	~1	38
HLAmismatch	147.7	~0	~1	40
Antigen	245.37	~0	~1	40
Alel	245.37	~0	~1	40
HLAgrI	245.37	~0	~1	40
Recipientage10	244.42	211.71	0.13	249.71
Relapse	151.19	116.22	0.23	154.22
aGvHDIIIIV	184.12	107.76	0.41	145.76

Table I : Showing the difference of deviances divided by the null deviance to identify the correlated variables.

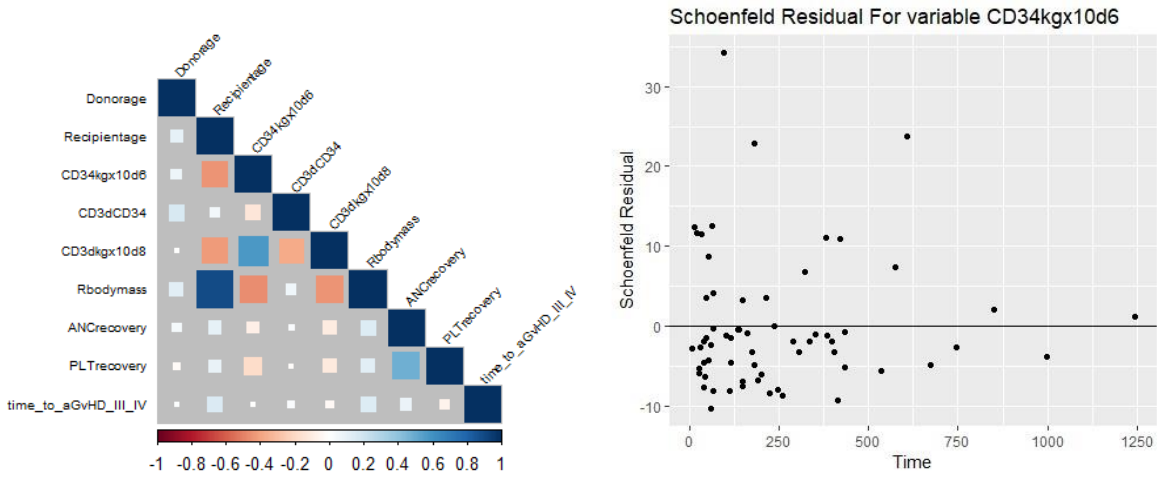


Figure 1 : (On the left) Matrix Correlation plot to show the correlation (in magnitude and direction both) among the continuous variables. Strong blue indicates a value of 1. Figure 2 : (On the right) Schoenfeld Residual plot for the variable CD34kgx10d6 - CD34+ cell dose per kg of recipient body weight (10^6/kg). Presence of clustered residuals. Time dependency for this variable exists.

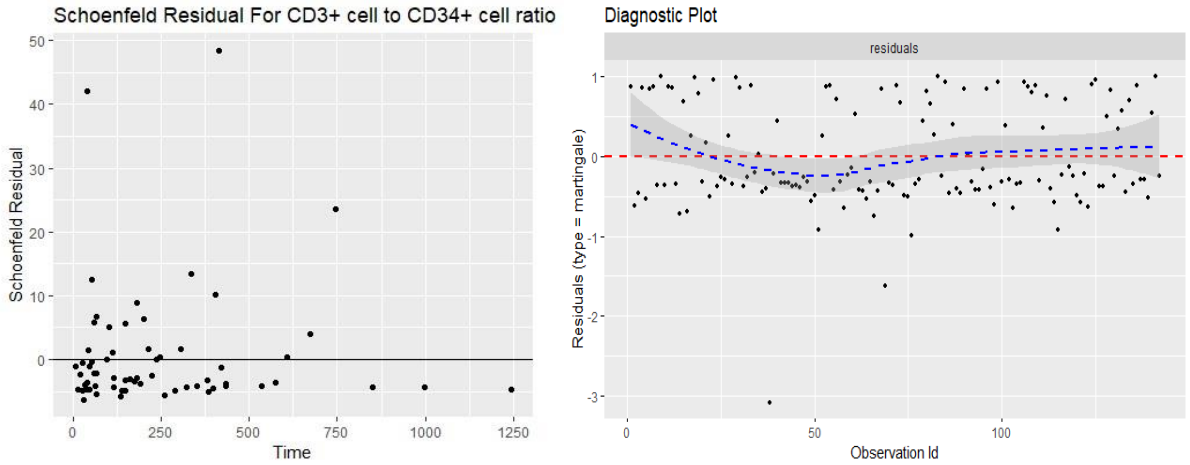


Figure 3 : (On the left)) Schoenfeld Residual plot for the variable CD3dCD34 - CD3+ cell to CD34+ cell ratio. Presence of clustered residuals. Time dependency for this variable exists. Figure 4: (On the right) Martingale Residual for the model with all the variables excluding collinear variables. Presence of outliers.

Table 2

Nested Models	LogLikelihood	AIC
Full Model (Excluding collinear variables)	69.68	564.48
Full Model (Excluding collinear variables) with Interaction	70.21	565.95
With Interaction term, excluding one of the cont cor variable	71.13	563.03
Without Interaction term, without correlated cont variable	69.47	562.69
Model after Step Regression	58.89	549.278

Table 2 : Showing the comparison among the nested models in terms of Loglikelihood and AIC

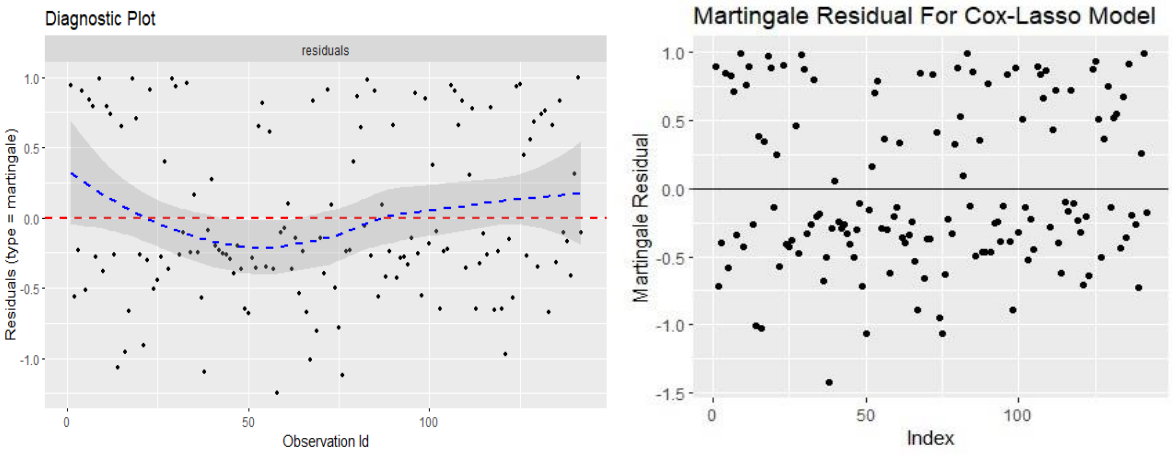


Figure 5 : (On the left) Martingale Residual plot for the model with no interaction and no correlated variable. The randomness is more evident and all the residuals are between -2 to 2 hence analogues to the classical residuals, we can exclude the idea of presence of outliers. Figure 6 : (On the right) Martingale Residual plot for the model fitted by LASSO regularization.

6.1.Healthcare data are valuable:

- Healthcare data are valuable and rarely available. It takes lots of money, man power and time to conduct researches, clinical trails and obtain data on medicine and epidemiology and hence, we might not want to throw away variables or observations which are information on participants or patients just because they are not statistically significant.

7. Cox-LASSO

- We want to retain as many as possible variables given that does not affect the model performance, and that is why we bring regularization into the picture. We perform the L1 penalization that is LASSO penalization, Cox-LASSO, for the Cox Regression.
- ***Dealing with Multicollinearity***:The good part is, we can involve all the variables, since LASSO penalization counts for as a fix for multicollinearity as well. It shall shrink the necessary coefficients to absolute zero.
- Here the ***choice of the tuning parameter*** *lambda* ***is based on the subjective choice instead of cross validation to serve our purpose***. A couple of tuning parameter and non zero coefficient dependency visuals have been given.
- ***We take the value of the tuning parameter to be 3 which results in maximum number of Non Zero variables (15). Based on the prior knowledge and subjective choice, 15 most significant variables to serve our purpose***.(Table 3 and 4, Figure 7,8,9 and 10)

8. Diagnostic Plot : Martingale Residual Plot

We plot martingale residuals for the models before and after step regression and also for the model fitted by the LASSO Regularization to illustrate the better performance based on the variable selection.In Martingale residual plots based on the model after step regression and LASSO penalization, randomness is more evident and all the residuals are between -2 to 2 hence analogues to the classical residuals, we can exclude the idea of presence of outliers.***Plots from the full model to the step regression to the LASSO, keep getting more refined***.(Figure 4, Figure 5 and Figure 6)

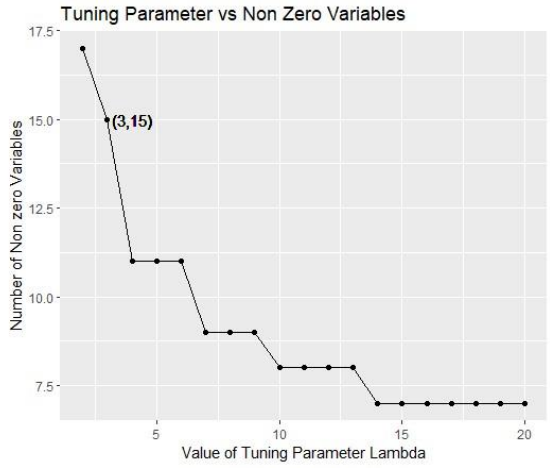
8.1. suspected Non-Linearity

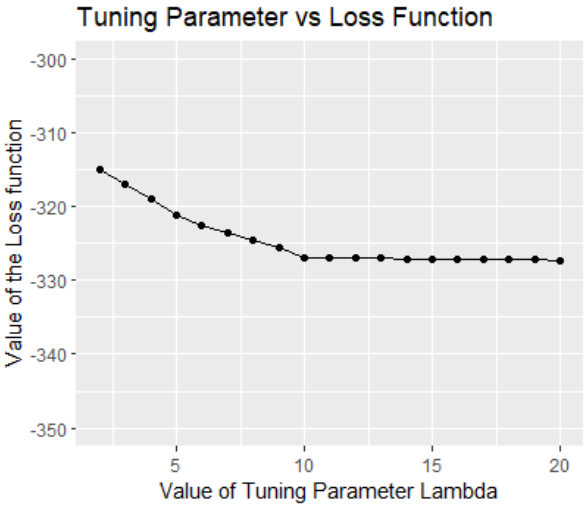
- ***From the Martingale Residual plot of the latest model, we observe a slight curvature*** and we suspect substantial presence of non linearity. Now, curvature could occur due to extreme value of certain variables. We could have inserted polynomial terms of suspected non linear variables. But that would bring the high chances of overfitting. Dealing with this situation with splines is a remarkable option but including splines would make the interpretability and computation less flexible.
- Curvature can also occur from variables having overall extreme values.
- Another option is to shift to some tree based algorithm. emanating from the various sectors, researchers are developing different algorithms using expert.The prediction performance of these algorithms is very important.An advantage of the tree based algorithm is that it does not require any transformation of the features if we are dealing with non-linear data because decision trees do not take multiple weighted combinations into account simultaneously.They are very fast and efficient compared to KNN and other classification algorithms.Tree based algorithms are non parametric.Easy to understand, interpret, visualize.Hence ***we introduce Random Survival Forest in our analysis***.

Table 3

LASSO model	No of Non-Zero Variables	Value Of Lambda	LOGLIKLIHOOD
Model I	17	2	-314.9293
Model 2	15	3	-316.9723
Model 3	11	4,5,6	-318.95, -321.116, -322.472
Model 4	9	7,8,9	-323.612, -324.525, -325.6238
Model 5	8	10,11,12,13	-326.85, -326.89, -326.94
Model 6	7	14-25	-327

Table 3 : showing the number of nonzero coefficients and loglikelihood of model based on the different values of lambda





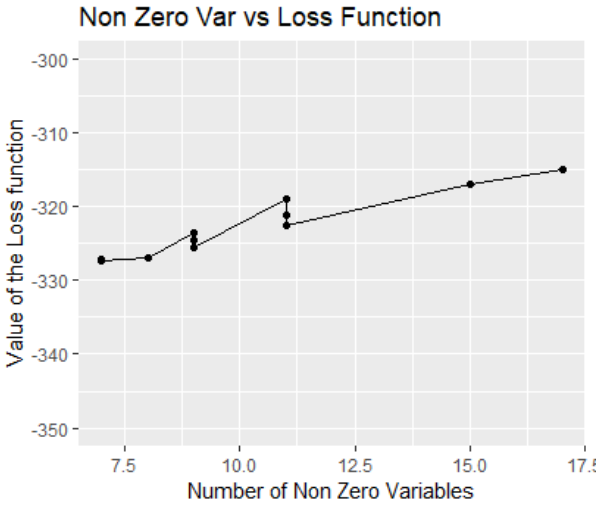


Figure 7 : (On the upper left) Increased value of the tuning parameter versus number of nonzero variables. Figure 8 : (On the upper right) A visualization on how the coefficients are getting reduced to absolute zero with increased value of the tunning parameter. Figure 9 : (On the lower left) Increased value of the tuning parameter versus Value of the loss function. Figure 10 : (On the lower right) Number of non zero variables versus Value of the loss function.

Table 4

Variable	Coefficient	Variable	Coefficient	Variable	Coefficient
Donorage	0.016809	Riskgroup1	0.013523	CD3dkgx10d8	-0.11062
IIIV1	0.054060	Recipientage	0.01224	Rbodymass	-2.055066e-03
DonorABO1	0.435295	Relapse1	0.88998	ANCrecovery	1.461394e-05
RecipientRh1	0.397243	CD34kgx10d6	0.13285	PLTrecovery	2.851586e-06
Disease	-0.01483	CD3dCD34	0.0010693	time_to_aGvHD	-6.389923e-07

Table 4 : showing the estimated coefficients of non zero variables given by Cox-LASSO model

Table 5

RSF Model	Splitting Rule	Variable tried at each Split	NodeSize	Error Rate	No of Trees used to train model
RSF1	Logrank Score	10	9	0.4052	5000
RSF2	Logrank random	10	9	0.3317	5000
RSF3	BS gradient	10	9	0.3325	5000
RSF4 with LASSO variables	Logrank random	5	9	0.3207	5000

Table 5 : showing comparisons among different RSF models with different splitting rules and variable considered at each split. Apparently, splitting rule Logrank random produces less Error rate for the data and if we pass only the variables selected by Cox-LASSO regression then the Error rate further decreases.

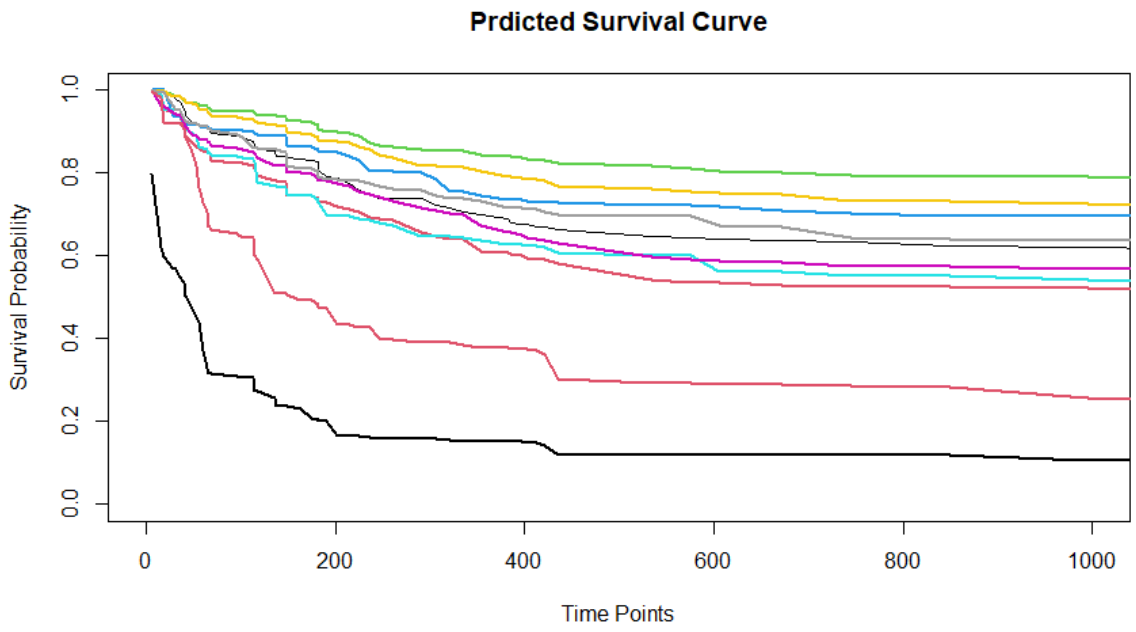


Figure 11: Predicted Survival Curve by RSF

9. Random Survival Forest (RSF)

9.1. Why you should shift to RSF:

- RSF can capture complex nonlinear relationships between predictors and survival outcomes. Unlike linear models like Cox regression, RSF does not assume proportional hazards or linear effects, allowing it to better capture the intricate relationships that may exist in the data.

- RSF can automatically handle variable selection by evaluating the importance of each predictor in the model. It can identify the most influential predictors for survival prediction, helping to reduce model complexity and improve interpretability.

- RSF is a nonparametric approach that does not make strict assumptions about the underlying distribution of the survival data. It can provide more flexible modeling and better capture complex survival patterns without imposing strong distributional assumptions.

- RSF can naturally handle time-varying effects and interactions, which is often a challenge in survival analysis. It can capture changing relationships between predictors and survival outcomes over time, making it suitable for analyzing dynamic survival data.

9.2. Our Analysis using RSF

Fitting a Random survival Forest model requires setting up a large number of parameters.

- Number of trees to train the model, Splitting rule, resampling method, if missing data imputation is needed nor not, nodesize, nodedepth etc and the preferred choices of these hyperparamters varies depending on the questions. Whether we focus on the variable importance or on the OOB error rate or anything apart from these.
  But here, ***our primary focus is to get as minimum as possible estimated error rate since it is a predictive setup***.

- We have tuned the node size for our data and selected the nodesize that would produce on an average less OOB error. which turned out to be 9.

- We fitted and compared different models using different splitting rules and different set of predictor variables. Trained the models with large number of trees(5000).

- For our analysis, logrank random split rule happened to produce less error while predicting a new value and if we pass the variables selected by LASSO with the logrank split rule the error rate further decreases.

- Prediction : we have passed the test data onto our model and plotted the predicted survival probability.(Figure 11)

- Now, we had suspected non linearity for certain variables, but curvature can occur in diagnostic plots for several reasons. And if there is no substantial amount of complexity and non linearity in the data, we should not expect jaw dropping performance from RSF models.

Discussion

- The analysis procedure highly depends on what kind of question(s) we want to answer. This project work is a predictive setup and not a mere jumble of machine learning algorithms. Our aim and concerns revolve around things like, feature selection, model accuracy and less predicting error on unseen data which all direct towards the improvement of the predictive power of our model.

- We started with our survival data(Bone Marrow data) and as an important part of the analysis, we perform data preprocessing** where we **fixed the issues regarding data imbalance by clubbing the levels of the categorical data and missing observations by deleting them due to complexity of them and Exploratory Data Analysis where we examine for the correlation and association among the variables.

- Most of the continuous variables are not even moderately correlated except for Body mass of the recipient and their age at the time of transplantation. *For the categorical variables, we tried to come up the measure* of the multiple association by fitting logistic regression and letting the other categorical variables play the role of predictors. The *difference of the null deviance and residual deviance divided by the null deviance(to scale them between 0 and 1) pictures the relationship among them as high value of such metric implies strong relation among them.*

- – As Diagnostic for CoxPH, we obtained schoenfeld and martingale residuals, for the continuous variables, the *schoenfeld residual showed patterns and clusters and to deal with suspected time dependency and violation of PH assumption, we brought stratified cox into the picture.*

- If we exclude all the collinear variables, after performing the step regression, we end up with 10 statistically significant variables, which does not serve our purpose. ***Healthcare data are rarely available and take a great deal of time and money to collect, so a clinician or an investigator might not want to throw away data or variables just because they are not statistically significant. We would always want to retain as many as possible variables given it does not affect the model optimization***.

- Here is where we part ways with the traditional and most common approaches of dealing with survival data. Since we want to keep as many as possible variables, we shall incorporate the penalization used in classical statistics as well as in machine learning setups. We shall incorporate L1 that is ***LASSO penalisation for the cox model***, known as Cox-LASSO. We can apply this on the overall data, no variable selection prior to this needed as for variables with lower importance, LASSO will shrink the parameter of the variable in question to absolute zero.

- The choice of lambda has always been a hot cake. Usually we perform cross validation to decide what the value of lambda should be. But in our situation, the cross validation was giving us a very small value of lambda which was potentially reducing the shrinkage power, hence for trying different values of lambda given the value are greater than 1, based on our subjective choice and prior knowledge, *we choose the value of lambda to be 3 which was resulting into 15 most subjectively significant variables and moderately less value of the loss function.*

What a clinician or an investigator is getting out the analysis and discussion so far

- For example, if two patients, say, A and B who are of same age, if A belongs to Riskgroup1 and B belongs to Riskgroup0, then compared to the baseline hazard, for person A for belonging to

riskgroup1, the hazard ratio that is the relative risk in comparison to person B, increases by 1.02.

- Also, if two patients, say, A and B who are of same age, belong to same riskgroups and have similar features, one significant unit increase in CD3+ cell dose per kg of recipient body weight, for person A, hazard ratio that is the relative risk in comparison to person B,decreases by 0.9

- Again, if two patients, say, A and B who are of same age, belong to same riskgroups and have similar features, if Reoccurrence of the disease has happened for patient B, then the hazard ratio that is the relative risk or the rate of happening the event in the next unit of time, in comparison to person A, increases by 2.4

- Next, we shift our focus back to the predictive setup, upon plotting the martingale residuals for the LASSO model, we *suspect a non linearity for the variables*.Now, curvature could occur due to extreme value of certain variables. We could have inserted polynomial terms of suspected non linear variables. But that would bring the high chances of overfitting. Curvature can also occur from variables having overall extreme values. To gain better performance, we turn to tree based algorithm. But a bunch trees, ***a forest, is always better than a single tree, hence we perform Random Survival Forest***, which is the extension of random forest based on survival setup splitting rule and is for right censored data.

- By Comparing different RSF models, upon setting the preferred values for different hyperparameter, for our data, ***we were able to come up with a model with error rate 0.32*** which is substantially less than 0.5(random guessing). That is, ***if we apply our RSF model to unseen data, the possibility of wrongly predicting the survival outcome is 32%. We achieved such an error by passing only the LASSO selected variables to the RSF model. This brings the future scopes of this study***.

Future Scope Of Studies

As discussed above, that among all the splitting rules, the logrank random rule resulted in the lowest error rate in comparison to others,but that error rate further decreased when instead of passing all the variables to the RSF model, we passed only the variables selected by LASSO penalization.
This leaves an open argument to whether we can ensemble these semi parametric, penalization and non parametric models to achieve even better accuracy and performance. RF performs, usually well, but for higher and much higher dimension it needs some level of regularization.
Hence the idea of Cox-LASSO-Forest could be very tempting, where we let the forest model know some part (could be linear) of the variability or the relations which boosts the performance of the forest model.
Due to computational limitation, we leave this tempting attempt to hang in the air.

REFERENCES

Reference

- Cox, D. R. "Regression Models and Life-Tables." Journal of the Royal Statistical Society. Series B (Methodological), vol. 34, no. 2, 1972, pp. 187–220. JSTOR, http://www.jstor.org/stable/2985181.
- Zhang, Hao Helen, and Wenbin Lu. "Adaptive Lasso for Cox's Proportional Hazards Model." Biometrika, vol. 94, no. 3, 2007, pp. 691–703. JSTOR, http://www.jstor.org/stable/20441405.
- Ishwaran, H., Kogalur, U.B., Blackstone, E.H. and Lauer, M.S., 2008. Random survival forests.
- Statistical Methods for Survival Analysis, John Wiley, Third Edition.
- Survival Analysis: Techniques for Censored and Truncated Data By John P. Klein, Melvin L. Moeschberge, 2nd Edition, Springer.
- Survival Analysis: A Self-Learning Text By David G. Kleinbaum, Mitchel Klei,2nd Edition, Springer.